DOCUMENT RESUME

ED 306 232 TM 012 988

AUTHOR Lunz, Mary E.; And Others
TITLE Variation among Examiners and Protocols on Oral Examinations.
SPONS AGENCY American Society of Clinical Pathologists, Chicago, Ill.
PUB DATE Mar 89
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Software; Difficulty Level; *Evaluators; *Examiners; Higher Education; Interrater Reliability; Latent Trait Theory; *Licensing Examinations (Professions); Medicine; Professional Education; Protocol Analysis; Rating Scales; Test Format; *Testing Problems; Test Items; Test Use; *Theory Practice Relationship
IDENTIFIERS *Facet Analysis; Medical Specialty Boards; *Oral Examinations; Rasch Model; Variability

ABSTRACT

A method for understanding and controlling the multiple facets of an oral examination (OE) or other judge-intermediated examination is presented and illustrated. This study focused on determining the extent to which the facets model (FM) analysis constructs meaningful variables for each facet of an OE involving protocols, examiners, and candidates; and establishing that the OE data fit the FM. The FM for calibrating all facets of an examination is an expansion of the Rasch model. The model was applied to data collected from an OE administered by a medical specialty board to 293 candidates. The same 48 examiners judged candidates on 18 protocols of case management, examining each candidate on three protocols. The FACETS computer program for the Rasch analysis of examinations with multiple facets was used to analyze the data. The FACETS program produces three statistics for each element of each facet: (1) a logit measure; (2) its standard error; and (3) the model fit statistic (infit and outfit). Results show that variables for protocol difficulty and examiner severity can be constructed. Examiners varied markedly in their level of severity, but were consistent in that level across candidates. Protocols and examiners generally fit the expectations of the model. When a protocol or examiner does not perform as expected, the fit statistic flags the situation. The FM applies a complex and powerful analysis to the examination situation that requires a series of human judgments combined with item writing methodology, rating scale definitions, and candidate interviews. Three data tables and one graph illustrate application of the model. (SLD)

ED306232

Variation Among Examiners and Protocols on Oral Examinations

Mary E. Lunz
American Society of Clinical Pathologists

John A. Stahl
American Society of Clinical Pathologists

Benjamin D. Wright
University of Chicago

John M. Linacre
University of Chicago

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARY E. LUNZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA., March, 1989

BEST COPY AVAILABLE

Variation Among Examiners and Protocols on Oral Examinations

Abstract

This paper discusses objectivity in judge-intermediated examinations. A method for calibrating all facets of an examination using the facets model, an expansion of the Rasch model, is presented and illustrated. Variables are constructed for candidates, protocols and examiners of an oral examination. The fit of the elements of each facet to the expectations of the model is reviewed. These data were found to be suitable for constructing measures. The implications of different combinations of examiners and protocols on the interpretation of candidate ability measures are discussed.

Key words: Oral examinations, psychometrics, IRT, judge severity, item difficulty.

Acknowledgment: This research was supported by, but does not necessarily reflect the policy or opinion of the American Society of Clinical Pathologists.

Variation Among Examiners and Protocols on Oral Examinations

INTRODUCTION

Oral examinations in medical education have been promoted as well suited to observe a candidate's ability to assemble clinically pertinent information systematically, test problem solving skills and observe communication skills and clinical judgment (Small, 1982). Oral examinations, in theory, provide the examiners with the opportunity to probe a candidate's reasoning, ethics and knowledge.

The validity and reliability of oral examinations, however, have been questioned because of subjectivity and potential bias (Hurley, 1982) related to examiners who are often found to be inconsistent. In an attempt to improve uniformity among examiners, structured protocols, standardized grading criteria and administration procedures, as well as examiner training, have been instigated for some oral examinations (Reed, 1982). This improves the similarity among examination materials and conditions, but serves only to direct the attention of the examiner, not to control the subjectivity of examiner assessments.

Inconsistency among judges has been studied extensively. Littlefield, et al (1981) compared the ratings of various types of judges (i.e. faculty and residents) and found significant differences in their assessments of similar student clerkships. Also, a multiple choice examination was found to be more reliable than the clinical ratings. Cason and Cason (1984) postulated that the ratings received by a subject are a function of the subject's true ability and the rater's characteristics including the rater's resolving power, sensitivity, stringency, effective rating floor and ceiling. A significant rater stringency effect and a significant student ability effect were found. de Gruijter (1984) demonstrated

differences among judges using linear and nonlinear analysis models. Lunz, et al (1988) found that judges demonstrate discernible levels of severity which affect examinee scores. The results of these studies support the premise that individuals who rate, judge or examine have unique perspectives which interact with the examination materials and the candidates' performances. Applying standardized grading criteria, protocols and administration procedures defines the examination process, but will not remove differences in judge perspective and severity. Each person who serves as a judge retains a unique perception of the examination materials, the standards for competence and the rating criteria.

The aim of a measurement system is to differentiate among candidates while generalizing across the particular variations in the other facets of the examination (Allal, 1988) such as protocols, examiners and the rating criteria. The studies in the literature often emphasize one facet such as examiner severity. A method for understanding and then controlling the multiple facets of an oral or other judge-intermediated examination is needed for reliable objective measurement.

A THEORY OF MEASUREMENT FOR EXAMINATIONS WITH MULTIPLE FACETS

It is anticipated that candidates have different levels of ability to perform the tasks represented on the examination. The goal is to measure the candidate's ability after accounting for the intermediating facets. The two facet (dichotomous) model used for analyzing multiple choice examinations includes facets for item difficulty and candidate ability. An examination with three or more facets, must include facets for candidate ability, item difficulty and any other intermediating facet, such as examiner severity, to analyze the process completely. The problem is how to measure candidate ability objectively when multiple facets are a

necessary part of the examination environment. As Jones (1981) explained, "measurement is only sensible when you know a good deal about what you are talking about." Psychometric analyses clarify our understanding of dichotomous measurement. It is possible that psychometric analyses can be expanded to facilitate our understanding of examinations with multiple facets.

When an examination with multiple facets is conceived, a series of assumptions, outlined by Kinston (1985), are incorporated into the belief system which gives the examination its value. The first assumption is that the examination has some degree of objectivity. It is assumed that the measurement of each candidate is independent of the protocols, examiners and rating scale used. If a candidate is found to be competent, this competence is intended to generalize beyond the particular sample of skills and knowledge included in the examination.

A second assumption is that distinctions can be made among candidates. The most able candidate will earn the highest score while less able candidates will earn lower scores regardless of the elements within each facet of the examination.

A third assumption is that examination items or protocols vary in difficulty. Examination developers know intuitively, if not explicitly, that the protocols presented vary in difficulty. Therefore, it is probable that a range of difficulty is represented across protocols although the psychometric tools to demonstrate this have been unavailable.

A fourth assumption is that value can be attached to the quality of a candidate's performance. Definitions of quality performance reside in the minds of the examination developers and oral examiners. Training oral examiners and establishing well-defined rating criteria have been thought to be a means of enhancing similarity. Inter-judge reliability studies,

however, often show variation beyond the expected error variance despite these efforts.

Finally, it is assumed that performance on the examination generalizes so that sensible action can ensue. A candidate who passes the examination is certified as having demonstrated an acceptable level of skill and knowledge, regardless of the specific sample of protocols or examiner characteristics challenged.

A measurement model designed to analyze an examination with multiple facets must provide an analysis of each facet of the examination. The particular elements within each facet must be accounted for so that a relevant probability statement about candidate performance can be constructed. The ordering of examinees must function as though it were independent of both the examiner and the items used by the examiner in assessing the performance of the candidate.

The Rasch model (1960/1980) estimates item difficulty and candidate ability, and expresses them on a common log-linear scale. The probability of person n with ability B achieving a correct response on item i with difficulty D is modeled as $\log(P_{ni}/1-P_{ni}) = (B_n - D_i)$. The probability of a correct response is a function of the difference between the ability of the person and the difficulty of the item (Wright and Stone, 1979).

The rating scale model is an expansion of the Rasch model to include an additional facet for the effect of a rating scale. The probability of person n with ability B achieving a rating step x with difficulty F on item i with difficulty D is modeled as $\log(P_{nix}/P_{nix-1}) = (B_n - D_i - F_x)$ (Wright and Masters, 1982).

The facets model is a further expansion of the Rasch model to include all facets which are pertinent to an examination, such as examiner severity. The probability of person n with ability B achieving rating step

x on item i with difficulty D from judge j with severity C is modeled as $\log(P_{nijx}/P_{nijx-1}) = (B_n - D_i - C_j - F_x)$ (Linacre, 1987). This facets model constructs a variable, for the elements within each facet of an examination, on the same log-linear (logit) scale. The ordering of the elements within each facet provides a frame of reference for verifying the intended examination definition. It becomes possible to observe candidate measures (B_n) from highest to lowest, protocols (D_i) from most to least difficult and examiners (C_j) from most to least severe. It is also possible to observe how the rating categories (F_x) are used by the examiners.

Candidate ability measures are calculated after the severity of each examiner and the difficulty of each protocol are calibrated and taken into account. This removes any bias introduced by the particular elements of the intermediating facets. It is still necessary to orient examiners to the examination materials and to the criteria for rating. But it is not necessary to depend on similarity among examiners or on the unlikely assumption that, because examiners are sampled, differences among them are random. The severity of each examiner and the difficulty of each protocol is located on the logit scale. Differences are expected and adjustments can be made.

OBJECTIVES

The objectives of this study are: (1) to determine the extent to which the facets model analysis constructs meaningful variables for each facet of an oral examination involving protocols, examiners and candidates and (2) to establish that these oral examination data fit the facets model.

METHODS

Data

Data were collected from the oral examination administered by a medical specialty board. Two-hundred-ninety-three candidates took the examination. Each candidate had two 20 minute oral examination sessions. Candidates were assigned to examiners who had no information about their education or experience. Each candidate was evaluated by two different examiners.

Eighteen structured protocols were constructed for the oral examination. Each protocol described the nature of the case. The candidate then acquired additional information from the examiner until a diagnosis could be made or a treatment determined. The protocols, for reasons of security, were divided into two groups for administration in the morning and afternoon. The distribution across content areas was comparable for the two groups of protocols.

The same 48 examiners judged the candidates in the morning and afternoon. This created an overlapping network of examiners with candidates and protocols. Prior to the examination, the protocols were reviewed by the examiners who agreed on the criteria for judging each protocol. Examiners were instructed to use three protocols in each session. Examiners judged candidates on each protocol on a four point scale (excellent = 3, acceptable = 2, marginal = 1, unacceptable = 0). Each candidate was meant to have 6 scores (3 protocols X 2 examiners); however, there was some variation because examiners did not always follow the directions.

Analysis

FACETS (Linacre, 1988) a computer program for the Rasch analysis of examinations with multiple facets is used to analyze these data. FACETS places each element within each facet on a common variable according to the following probability equation.

$$\log \left[\frac{P_{nijk}}{P_{nijk-1}} \right] = B_n - D_i - C_j - F_{ijk}$$

where P_{nijk} = probability of candidate n being given a rating of k by judge j on protocol i

P_{nijk-1} = probability of candidate n being given a rating of k-1 by judge j protocol i

B_n = ability of candidate n

D_i = difficulty of protocol i

C_j = severity of judge j

F_{ijk} = threshold difficulty of being rated in category k on protocol i, relative to category k-1 by judge j

Three facets are defined in this oral examination.

Facet #1: (candidate ability, B_n) is the ability measures for each of the 293 candidates. The logit ability measures are corrected for examiner severity and protocol difficulty so that the bias of the particular examiners or protocols is mitigated. The candidate ability measures are transformed to scaled scores.

Facet #2: (protocol difficulty D_i) is the difficulty calibrations for each of the 18 protocols. These protocols are divided into 2 groups for morning and afternoon administration. The distribution of protocols across content areas was comparable for morning and afternoon. In the analysis, each group of 9 protocols is centered at zero so that any difference in the range of their difficulty can be observed.

Facet #3: (examiner severity, C_j) is the severity calibrations for each of the 48 examiners. Examiner severity calibrations are calculated across protocols and candidates.

The FACETS program produces three statistics for each element of each facet, a logit measure, its standard error and model fit statistic (infit and outfit). The purpose of the fit statistics is to ascertain the suitability of these particular data for constructing variables and making measures (Wright and Stone, 1979). The fit statistics for examiners

indicate the degree to which each examiner's ordering of candidates is consistent with the estimated candidate ability measures. The fit statistics for protocols indicate the degree to which examiners apply consistently their criteria for quality. The infit is the mean-squared residual across all cases sensitive to deviations at the point where decisions are being made. The expected value is 1.00. The outfit is the mean-squared residual across all cases sensitive to outliers which show as unexpectedly high or low ratings. The expected value is again 1.00. The criteria for flagging deviations from expected is set at less than or equal to 0.50, or greater than or equal to 1.50. Means, standard deviations, a separation reliability and chi-squares for homogeneity are also calculated for each facet.

RESULTS

The first objective is to determine the extent to which the facets model constructs meaningful variables for each facet of this oral examination. The second objective is to evaluate the extent to which the oral examination data fit the facets model. The information related to these objectives is reported in Tables 1 and 2.

Table 1 lists the difficulty calibrations, standard errors, mean squared infit and outfit for each protocol. Each group of protocols is centered at zero. The protocols used in the afternoon have a slightly wider range of difficulty (1.81 to -1.24 logits) than those used in the morning (1.11 to -0.87 logits). The standard errors are consistent across protocols. Protocol 18 was used least frequently and has the largest standard error.

An examiner intermediates between the candidate and the protocol. Thus, the protocol fit statistics detect inconsistent ratings within

protocols across examiners. In this oral examination, all protocols fit to the expectations of the model. Examiners applied consistently their criteria for quality within protocols across candidates.

The separation reliability for protocols (comparable to a KR-20) is 0.64, indicating that the differences in difficulty of these protocols is discernible beyond error variance. This is supported by a highly significant chi-square of homogeneity ($p<.000$). These results confirm that protocols can be placed along a continuum of difficulty. The validity of the protocol difficulty calibrations was confirmed by discussions with the examination developers who explained why some protocols were more difficult than others.

Insert Table 1 About Here

Table 2 gives the examiners in severity order. Examiner severity calibrations include all protocols and candidates judged by the examiner. Examiners vary in severity from 1.08 to -2.00 logits demonstrating that some examiners are more severe than others. The standard errors of the examiner severity estimates are relatively consistent. Most examiners fit the model expectations. However, examiners 16, 24, 32, and 33 show a greater than expected degree of consistency (fit = .50 or less). They appear to be less sensitive to variations in protocol difficulty and candidate ability than the other examiners. Examiners 3, 12, 14, 34, and 41 show some inconsistency (fit = 1.50 or more). These examiners gave some unexpectedly high ratings to less able candidates and/or some unexpectedly low ratings to more able candidates. Further study of the scoring patterns

of these nine examiners is in order. The other examiners judged the protocols consistently and fit the expectations of the model.

The separation reliability across examiners is 0.82 and the chi-square for homogeneity is highly significant ($p<.000$). This confirms that the variation among examiners did not occur by chance, but that each examiner has a discernible level of severity, applied consistently in the case of most examiners, across all protocols and candidates. These results demonstrate that the examiners can be placed on a continuum for severity.

Insert Table 2 About Here

The calibrations for protocols and examiners indicate that the interpretation of candidate performance is likely to be biased by the particular examiners and protocols used, unless a correction for the effects of these intermediating facets is introduced before the candidate ability measures are calculated. A convenient way of analyzing the combined effect of protocol difficulty and examiner severity on candidate ability is to review the configuration of facet elements associated with candidates that received the same raw score.

Table 3 reports six selected candidates with varying combinations of examiners and protocols, all of whom received a raw score of 12. If only the raw score were considered, these candidates would be viewed as equivalent. Consideration of the severity of the examiners and difficulty of the protocols should produce a more precise ability measure for these candidates and demonstrate that a single raw score can represent a range of candidate ability given the particular elements of the intermediating facets.

Insert Table 3 About Here

Candidate 1 achieved the lowest ability measure. The examiners that judged this candidate are lenient (-1.45 logits) and the mean difficulty of the protocols is -0.09 logits. When this combination of very lenient examiners and fairly easy protocols is accounted for, a lower candidate ability measure results.

Candidate 2 was judged by more severe examiners. The mean difficulty of the protocols is 0.24 logits and the mean severity of the examiners is -0.51 logits. Candidate 2 had to demonstrate more ability than did candidate 1 to achieve the same raw score because he/she had more difficult protocols and more severe examiners. The candidate measures reflect the difference in ability between candidates 1 and 2 because the effect of the intermediating facets is taken into account.

Candidate 3 was judged by more severe examiners (-0.10 logits) than candidates 1 and 2. However, the mean difficulty of the protocols is only -0.16 logits. Given the combination of more severe examiners and easier protocols, candidate 3 demonstrates a relatively comparable ability measure to that of candidate 2.

Candidates 4 and 5 were judged using protocols of comparable average difficulty (0.11 logits). The average severity of the examiners is virtually identical (.29 and .30 logits). In achieving a raw score of 12, these two candidates demonstrate the same ability under very similar examination conditions. This is the ideal examination situation and shows that the facets model analysis produces the same results as raw scores when

the examiners and protocols are comparable for candidates of similar ability. The examiners of candidates 4 and 5, however, were more severe and the protocols more difficult than those of candidates 1, 2 and 3. When this is accounted for, candidates 4 and 5 earn higher ability measures.

The protocols presented to candidate 6 are relatively difficult (0.17 logits) and the examiners are quite severe (0.97 logits). Candidate 6 had to demonstrate a higher level of ability in order to fulfill the expectations for quality of these examiners on these particular protocols. When the correction for protocol difficulty and examiner severity is made, candidate 6 is shown to be significantly more able than candidate 1 who also earned a raw score of 12 from lenient examiners on easier protocols.

Graph 1 gives the spread of ability measures across raw scores. These data show that candidates who receive identical raw scores, which do not account for differences in protocols and examiners, often have different levels of ability when these intermediating factors are taken into account. It is only after accounting for the differences in examiner severity and protocol difficulty that unbiased estimates of candidate ability can be achieved.

INSERT GRAPH 1 ABOUT HERE

DISCUSSION

The results of this study demonstrate that variables for protocol difficulty and examiner severity can be constructed. Examiners vary markedly in their level of severity, but tend to be consistent in that level of severity across candidates. Protocols and examiners generally fit

the expectations of the model. When a protocol or examiner does not perform as expected, the fit statistic flags it. Additional training of examiners or revision of protocols may then be appropriate.

The FACETS analyses statistically define the variable for each examination facet and also mitigate the bias of the particular elements of any facet before calculating candidate ability measures. Examiner bias cannot be eliminated, but it can be accounted for using these techniques so the examination results meet the requirements for objectivity (Wright & Douglas, 1986) and fairness.

The facets model applies a complex and powerful psychometric analysis to a complex examination situation which requires a series of human judgments combined with item writing methodology, rating scale definitions, and candidate interviews. When less sophisticated techniques such as "handicapping a judge" are applied, only one of the multiple facets which combine to create the examination situation, is considered. This serves to weight that facet disproportionately, or to create an interdependence among facets which cannot be sorted out. Methods which appear to be simpler, can further confuse, rather than clarify, the interpretation of candidate performance.

For multiple choice examinations, all facets are accounted for in item difficulty and person ability. For oral and other judge intermediated (i.e., essays and practicals) examinations, all facets must be taken into consideration. This is the only way to insure objectivity and acquire a full understanding of the examination process.

References

Allal, L. (1988) Generlizability theory in J.P. Keeves (ed.) Educational Research, Methodology and Measurement: An International Handbook, New York: Pergamon Press.

Cason, G.J. and Cason, C.L. (1984) A deterministic theory of clinical performance rating. Evaluation and the Health Professions, 7, 221-247.

de Gruijter, D.N.M. (1984) Two simple models for rater effects. Applied Psychological Measurement, 8, 2, 213-218.

Hurley, H.J. (1982) The Part II Certifying Examination in Dermatology: An Assessment of Interpretive Skills in J.S. Lloyd (ed.), Evaluation of Noncognitive Skills and Clinical Performance; Chicago, IL: American Board of Medical Specialties.

Jones, R.V. (1981) Some consequences of physics. Nature, 293, 23-25.

Kingston, W. (1985) Measurement and the structure of scientific analysis. Systems Research 2, 95-104.

Linacre, J.M. (1987) An Extension of the Rasch Model to Multi-faceted Situations, Chicago: University of Chicago.

Linacre, J.M. (1988) FACETS, Chicago, IL; Mesa Press.

Littlefield, J.H., Harrington, J.T., Anthracite, N.E., Garmon, R.E. (1981) A description and four year analysis of a clinical clerkship evaluation system. Journal of Medical Education, 56,334-340.

Lunz, M.E., Linacre, M., Wright, B. (1988) Impact of Judge Severity on Examination Scores. Paper presented at American Educational Research Association, New Orleans, LA.

Rasch, G. (1960/1980) Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press.

Reed, G.F. (1982) The Use of Visual Materials in the Evaluation of Interpretive Skills in J.S. Lloyd (ed.), Evaluation of Noncognitive Skills and Clinical Performance; Chicago, IL: American Board of Medical Specialties.

Small, S.M. (1982) Evaluation Methodology for the Oral Examination of the American Board of Psychiatry and Neurology in J.S. Lloyd (ed.), Evaluation of Noncognitive Skills and Clinical Performance; Chicago, IL: American Board of Medical Specialties.

Wright, B., Stone, M. (1979) Best Test Design, Chicago, IL. Mesa Press.

Wright, B., Masters, G. (1982) Rating Scale Analysis, Chicago, IL Mesa Press.

Wright, B., Douglas, G. (1986) The Rating Scale Model for Objective Measurement, Chicago: Mesa Press. Memorandum No. 35.

Table 1

Calibrated Difficulties of Protocols

	Num	Measure Calibr. Logit	Model Error	Consistency Infit MnSq	Outfit MnSq
	Morning Protocols				
DIFFICULT	3	1.11	0.15	1.1	1.0
	2	0.63	0.13	1.2	1.1
	7	0.17	0.13	1.2	1.2
	4	0.07	0.24	0.7	0.9
	5	-0.10	0.14	0.9	0.9
	6	-0.14	0.14	1.2	1.1
	8	-0.23	0.15	0.9	0.9
	9	-0.65	0.22	1.1	1.1
EASY	1	-0.87	0.21	0.7	0.9
	Afternoon Protocols				
DIFFICULT	18	1.81	0.38	1.4	1.3
	15	1.06	0.16	1.2	1.1
	11	0.32	0.21	1.0	0.9
	16	0.11	0.20	0.9	1.0
	14	-0.04	0.19	0.7	0.7
	17	-0.20	0.19	0.7	0.7
	12	-0.75	0.19	1.1	1.0
	10	-1.06	0.18	1.1	0.9
EASY	13	-1.24	0.24	0.7	0.8
	Mean:	0.00	0.36	0.9	0.9
	S.D.:	0.78	0.30	0.3	0.3

Separation Reliability 0.64
chi-square of homogeneity: 270.61
d.f.: 27 significance:0.00

Table 2

Calibrated Severities of Examiners

		Measure		Consistency	
		Calib.	Model	Infit	Outfit
	Nu	Logit	Error	MnSq	MnSq
SEVERE	47	1.08	0.33	0.9	1.0
	16	1.07	0.24	0.5	0.4
	19	1.02	0.24	1.4	1.4
	40	0.89	0.29	1.2	1.1
	29	0.87	0.23	0.7	0.7
	33	0.76	0.26	0.4	0.5
	34	0.74	0.25	1.5	1.4
	36	0.65	0.36	0.9	0.9
	7	0.57	0.22	0.6	0.6
	15	0.55	0.26	1.2	1.2
	24	0.51	0.26	0.5	0.5
	43	0.49	0.31	0.8	0.7
	18	0.46	0.27	1.1	1.4
	13	0.45	0.24	0.8	0.7
	41	0.37	0.31	1.5	1.8
	14	0.31	0.25	1.5	1.3
	5	0.28	0.25	1.0	1.0
	6	0.27	0.31	1.2	1.1
	30	0.16	0.27	1.1	1.2
	38	0.15	0.27	0.7	0.7
	27	0.13	0.27	1.2	1.2
	4	0.12	0.25	0.8	0.7
	31	0.11	0.23	0.9	1.0
	37	0.07	0.27	1.0	1.0
	42	0.07	0.33	1.3	0.9
	8	0.04	0.27	0.7	0.7
	3	-0.04	0.30	1.6	1.5
	11	-0.06	0.27	0.6	0.6
	20	-0.10	0.30	0.9	0.6
	46	-0.11	0.30	0.8	0.7
	1	-0.14	0.24	0.8	0.9
	28	-0.16	0.29	1.1	1.0
	12	-0.17	0.27	1.5	1.5
	10	-0.18	0.25	1.2	1.2
	32	-0.19	0.31	0.5	0.5
	48	-0.24	0.29	1.4	1.3
	26	-0.32	0.27	1.4	1.5
	35	-0.41	0.25	0.6	0.5
	39	-0.68	0.30	1.0	1.0
	45	-0.73	0.25	1.4	1.2
	44	-0.75	0.26	0.7	0.7
	21	-0.85	0.31	1.1	1.0
	23	-0.86	0.34	1.0	1.0
	9	-0.90	0.28	1.2	1.0
	22	-0.98	0.30	1.1	1.0
	25	-1.01	0.33	1.4	1.2
	2	-1.34	0.32	1.0	1.1
LENIENT	17	-2.00	0.30	1.0	1.1
	Mean:	0.00	0.28	1.0	1.0
	S.D.:	0.66	0.03	0.3	0.3

Separation Reliability 0.82
chi-square of homogeneity: 255.57
d.f.: 47 significance:0.00

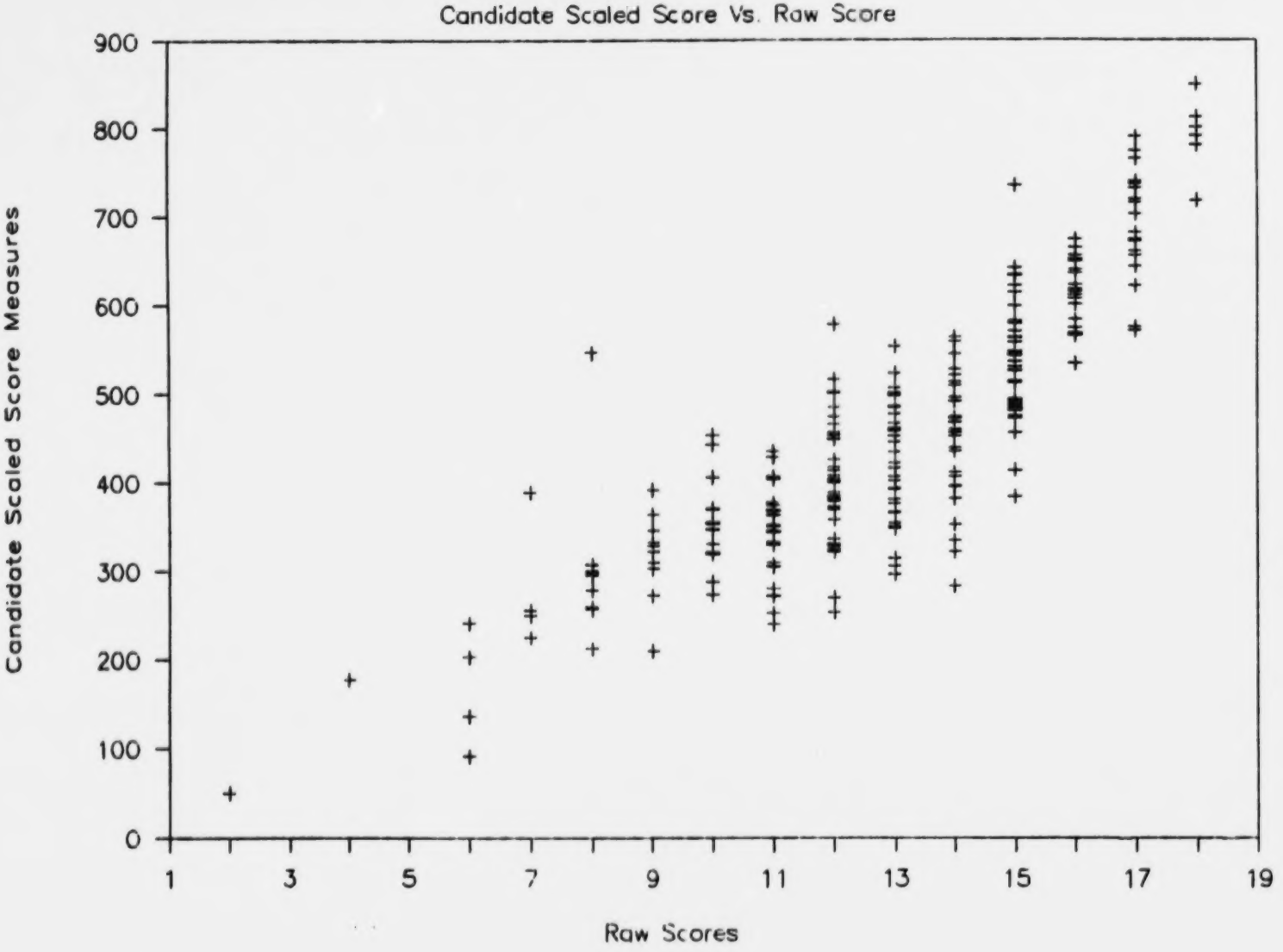
Table 3

Analysis of Selected Candidates who
Received a Raw Score of 12
(6 Protocols and 2 Examiners per Candidate)

Candidate Number	1	2	3	4	5	6
Candidate Scaled Score Measure*	254	381	379	449	450	516
Scaled Score Measure SE	66	68	67	67	67	54
Examiner Numbers	17/09	21/10	12/03	08/15	38/18	29/47
Mean Examiner Severity**	-1.45	-.51	-.10	.29	.30	.97
Mean Protocol Difficulty**	-.09	.24	-.16	.11	.11	.17

*Candidate measures and measure SEs are transformed to scaled scores.
Scaled score = ((100 * candidate logit measure) + 500)
**Examiner severity and protocol difficulty are reported in logits

Graph 1


Candidate Scaled Score Vs. Raw Score
Candidate Scaled Score Measures
900
800
700
600
500
400
300
200
100
0
1
3
5
7
9
11
13
15
17
19
Raw Scores